cap5771sp25
March 19, 2025

Exam 1 part 2 version a

Question 1. According to Data Wrangling lecture slides (Slide 32), what is the curse of dimensionality, and how does it impact machine learning models?

(a) `The curse of dimensionality is when datasets become too large to fit into memory, causing slow computations.`

(b) `More dimensions always improve model accuracy by providing additional features for learning.`

(c) `The curse of dimensionality only affects text-based datasets, not numerical or categorical data.`

(d) `As the number of dimensions increases, data points become more sparse, making clustering and distance-based models less meaningful.`

(e) `Increasing the number of dimensions makes feature selection unnecessary since all features contribute equally.`

Question 2. According to the Data Wrangling v2 lecture, which of the following is not true of Correlation Analysis?

(a) `Correlation analysis is one of the methods that can be used to identify redundant attributes in datasets.`

(b) `The smaller the value given by a Chi-Square test the greater the correlation.`

(c) `If our correlation analysis shows a high correlation between attributes, then it also implies causation between these attributes.`

(d) `Pearson's product moment coefficient can be used for correlation analysis of numeric data.`

(e) `The Chi-Square test is used for correlation analysis of nominal data.`

Question 3. According to the data wrangling lecture, which of the following statements best describes the method of binning for data smoothing?

(a) `It identifies outliers by calculating the Interquartile Range (IQR) and removing data points that fall outside 1.5 times the IQR from the quartiles.`

(b) `It involves partitioning sorted data into equal-frequency or equal-width bins and then smoothing by bin means or boundaries.`

(c) `It transforms data by scaling values to fall within a smaller, specified range, such as 0.0 to 1.0.`

(d) `It uses regression models to predict and replace noisy data values based on the relationship with auxiliary variables.`

(e) `It applies clustering algorithms to group similar data points and uses cluster centroids to represent and smooth data.`

Question 4. Based on the lectures on Data Wrangling, answer the following. Consider a dataset of temperature readings from various machines/sensors collected at varying time intervals. Few of the machines log data every second, while few does it every minute. What is the best approach to preprocess this data in order to ensure consistency in the dataset?

(a) `Duplicate the values for machines with less frequent recordings so that we can match all data to the highest logging frequency.`

(b) `Aggregate the readings over a fixed time window, like considering average temperature per minute, in order to standardize timestamps.`

(c) `Remove the temperature readings from machines that do not log at the same frequency , and maintain uniformity.`

(d) `Use interpolation methods to estimate/generate the missing values for the less frequent recordings.`

(e) `First normalize the dataset, and then adjust all the frequencies to match the highest recorded value.`

Question 5. According to slide 15 of the Data Wrangling Part 2 lecture, how can clustering be used to help identify outliers?

(a) `Clustering puts outliers in the largest cluster, since clusters are always part of the majority group.`

(b) `Clustering reduces the data set to only the most central data points, eliminating outliers.`

(c) `Clusters create distinct groups, so any points outside of those groups can be easily identified as a potential outlier.`

(d) `Outliers do not affect the clustering process and clusters cannot identify them.`

(e) `Clustering looks for data points that are close to the average value of the data set, revealing outliers in the process.`

Question 6. According to the lecture on the Data Wrangling Demo, what does this line of code do in descriptive_utils.py?
null_counts_rows[null_counts_rows == max_nulls].index.tolist()

(a) `It selects the row indices where the total count of missing values is greater than or equal to the maximum number of missing values found in the dataset.`

(b) `It selects the row indices where the number of missing values is equal to the minimum number of missing values found in the dataset.`

(c) `It selects the row indices where the number of missing values is equal to the maximum number of missing values found in the dataset.`

(d) `It selects the row indices where the number of missing values is equal to the average number of missing values across all rows.`

(e) `It selects the row indices where at least one column has missing values in the dataset.`

Question 7. According to the data integration part in data wrangling lecture slides(18-24), a national healthcare network is merging patient records from two independent hospital systems into a unified database. One system, used by Hospital A, stores Social Security Numbers (SSNs) in the format ”XXX-XX-XXXX”, while the other system, used by Hospital B, records SSNs as a continuous 9-digit number (”XXXXXXXXX”).

After integration, duplicate patient records emerge, insurance claims are mismatched, and critical medical histories are fragmented across multiple profiles. Additionally, some patients appear to have two separate billing accounts, while others become unidentifiable due to inconsistencies in the identification process.

What is the primary data integration issue here?

(a) `The same patient data is duplicated in different formats, causing unnecessary storage.`

(b) `Some patient records reference SSNs that do not exist in the integrated database.`

(c) `The SSN field is formatted differently across databases, requiring transformation.`

(d) `The same SSN has been recorded with inconsistent values, leading to identity mismatches.`

(e) `The same patient exists in both systems but is not recognized as a single entity.`

Question 8. According to the data wrangling demo, why is it recommended to handle outliers before imputing missing data?

(a) `To reduce computational overhead during the imputation process.`

(b) `To prioritize outlier removal over addressing missing data.`

(c) `To avoid introducing new outliers during the imputation process.`

(d) `Because outliers cannot be identified after missing data is imputed.`

(e) `To ensure that summary statistics like the mean or median are not skewed by extreme values.`

Question 9. A dataset contains student test scores: 50, 55, 60, 62, 65, 68, 72, 75, 80, 150. Using the Interquartile Range (IQR) method, which of the following values would be classified as an outlier?

(a) `No outliers exist in this dataset.`

(b) `50`

(c) `150`

(d) `80`

(e) `72`

Question 10. From the Data Wrangling Demo,

```
import pandas as pd
from scipy.stats import zscore
def correcting_datatypes(df, date_cols=None, categorical_cols=None, float_cols=None):
    if date_cols:
        for col in date_cols:
           if col in df.columns:
          try:
            df[col] = pd.to_datetime(df[col], format=“mixed”, errors=“coerce”)
            print(f‘Converted ’col’ to datetime’)
          except ValueError as e:
            print(f‘Warning: Could not convert column ’col’ to datetime. Error: e’)
          else:
             print(f‘Warning: Column col does not exist in the DataFrame.’)
```

In the function correcting_datatypes, we see the code snippet:

```
    df[col] = pd.to_datetime(df[col], format=‘mixed’, errors=‘coerce’)
```

What is the significance of setting format=‘mixed’ when converting a column to a datetime in this context?

(a) It automatically detects and corrects all invalid or incomplete date values without any user intervention.

(b) It ensures that dates are strictly interpreted with a single format, causing any values that don't match to be rejected.

(c) It indicates that the column may contain dates in multiple formats, prompting Pandas to attempt parsing each entry flexibly.

(d) It forces the parser to treat every value as UTC time, regardless of the original timezone or date string.

Question 11. According to the lecture materials, which method of data cleaning would be MOST appropriate when dealing with a dataset where values are expected to form distinct groups and deviations from these groups are considered errors?

(a) Linear regression smoothing

(b) Chi-square analysis

(c) Binning with equal-width partitioning

(d) Clustering-based smoothing

(e) Z-score normalization

Question 12. Based on the lecture and readings on Data Wrangling, what is the primary difference between incomplete data and intentional missing data in a real-world dataset?

(a) Intentional missing data can be ignored, whereas incomplete data must always be corrected.

(b) Incomplete data occurs due to unintentional errors or missing records, whereas intentional missing data is deliberately left blank or replaced with a default value.

(c) Incomplete data is always caused by privacy concerns, while intentional missing data occurs due to data collection issues.

(d) Incomplete data is always incorrect, while intentional missing data is always correct.

(e) Extra data existing constitutes as incomplete data, whereas missing intentional data results from the deletion of data.

Question 13. According to lecture on data mining principles, which of the following is NOT a data reduction strategy?

(a) `Numerosity reduction`

(b) `Dimensionality reduction`

(c) `Data compression`

(d) `Data migration`

(e) `Feature selection`

Question 14. According to Data Wrangling v2 slide 4, which Data Preprocessing task is applied when a university categorizes student enrollment data by class level (Freshman, Sophomore, Junior, Senior) and then further groups it into Undergraduate and Graduate programs for reporting purposes?

(a) `Data Cleaning.`

(b) `Concept Hierarchy Generation.`

(c) `Data Compression.`

(d) `Dimensionality Reduction.`

(e) `Data Integration.`

Question 15. According to the lecture Data Wrangling Part 2 (Slide 8), why is data cleaning considered an iterative process?

(a) `Data cleaning is a linear process that follows a strict set of predefined steps without revisiting previous stages.`

(b) `Data cleaning primarily focuses on formatting inconsistencies rather than detecting outliers or resolving missing values.`

(c) `Data cleaning involves continuous improvement by identifying and correcting errors over time.`

(d) `Once a dataset is cleaned, further cleaning is unnecessary unless new data is added .`

(e) `Data cleaning aims to eliminate all variations in the dataset, ensuring that every value is uniform.`

Question 16. According to the lecture on data wrangling part2 (slide no 21), Which of the following best describes the primary difference between data warehousing and virtual data integration?

(a) `Data warehousing physically consolidates data, whereas virtual integration accesses data in real-time without replication.`

(b) `Virtual integration requires all data sources to follow the same schema, while data warehousing does not.`

(c) `Data warehousing allows for real-time querying of distributed data, while virtual integration requires batch processing.`

(d) `Data warehousing is only suitable for structured data, while virtual integration supports both structured and unstructured data.`

(e) `Virtual integration stores data permanently, while data warehousing deletes data after each query.`

Question 17. According to the lecture on data wrangling demo, If null_counts_rows is a Pandas Series containing the number of null values in each row, what does the following command return?
null_counts_rows[null_counts_rows == max_nulls]

(a) `A list of row indices where the number of null values is equal to max_nulls.`

(b) `A filtered Series containing only the rows where the null count equals max_nulls.`

(c) `A NumPy array of null counts where they equal max_nulls.`

(d) `A boolean mask indicating which rows have exactly max_nulls null values.`

(e) `A count of how many rows contain max_nulls null values.`

Question 18. According to data-wrangling-v2 page 17. Which of the following processes is often used to combine data from different sources (virtual or actual) and provide users with a unified view of the data?

(a) `Feature Selection.`

(b) `Data Augmentation.`

(c) `Data Integration.`

(d) `Regression.`

(e) `Clustering.`

Question 19. According to Data wrangling lecture, In a machine learning project, two different techniques were used for attribute construction:

Combining features: Created a new feature by merging "purchase frequency" and "average transaction value" to better capture customer spending behavior.

Data discretization: Transformed a continuous "age" variable into categorical bins such as "young," "middle-aged," and "senior."

Considering these techniques, how do combining features and data discretization differ in their impact on a machine learning model?

(a) `Combining features primarily reduces dataset size, while data discretization improves model accuracy by increasing feature granularity.`

(b) `Both techniques are exclusively used for reducing overfitting in machine learning models.`

(c) `Both combining features and data discretization primarily serve the same purpose: reducing redundancy in the dataset.`

(d) `Combining features enhances predictive power by creating richer representations, while data discretization simplifies models and improves interpretability.`

(e) `Data discretization creates more detailed numerical features, while combining features simplifies the dataset.`

Question 20. According to data-wrangling-v2 slide 25 and associated lectures, which of the following can be determined by a chi-square test in data wrangling?

(a) `The mean of a type of numeric data`

(b) `Correlation in numeric data`

(c) `Causal relations in nominal data`

(d) `Causal relations in numeric data`

(e) `Correlation in nominal data`

Question 21. According to Data Wrangling v2, slide 38, which of the following is a key characteristic of parametric data reduction methods?

(a) `They do not require any predefined assumptions about data distribution.`

(b) `They rely exclusively on clustering techniques to reduce dimensionality.`

(c) `They always result in data loss due to lossy compression techniques.`

(d) `They store full datasets while reducing redundancy through compression.`

(e) They assume a specific mathematical model to approximate data.

Question 22. According to Data Wrangling lecture, what is a key difference between lossless and lossy compression?

(a) Lossy compression always results in better model performance.

(b) Lossy compression can reconstruct data with 100% accuracy.

(c) Lossy compression is always preferred in data science applications.

(d) Lossless compression preserves all original data, while lossy compression removes some information.

(e) Lossless compression reduces data size by removing outliers.

Question 23. A dataset contains salary information recorded in different currencies (e.g., USD, EUR). How would you standardize this data?

(a) Convert all salaries into a single currency using an exchange rate table before analysis.

(b) Normalize salary values between 0 and 1 using Min-Max scaling.

(c) Remove all salary records that are not in USD for simplicity.

(d) Group salaries by currency type without converting them.

Question 24. According to the lecture slides data-wrangling-v2, which of the following BEST describes the concept of “Independence of Semantic Variations” in the context of data integration?

(a) The ability of the system to operate with any data structure or syntax, handling databases like SQL or NoSQL.

(b) The ability of the system to recognize and reconcile differences in data meanings and usage across systems, such as understanding "DOB" and "DateOfBirth" as equivalent terms.

(c) The process of physically moving all data into a single data warehouse.

(d) The ability of the system to access data regardless of its physical storage location.

(e) The use of correlation analysis to identify redundant attributes across different databases.

Question 25. According to the lecture and readings on Data Wrangling, what is the key difference between metadata-based detection and rule-based detection in identifying data discrepancies?

(a) Metadata-based detection utilizes stored data attributes such as domain, range, and dependencies, whereas rule-based detection enforces predefined conditions like numerical ranges or logical constraints.

(b) Metadata-based detection modifies data directly, while rule-based detection only flags errors without modification.

(c) Rule based detection focuses on discerning quantitative information while metadata detection goes beyond numbers.

(d) Metadata-based detection requires AI models to function, while rule-based detection is manually performed.

(e) In the context of metadata-based detection the only capable action is finding query duplicate records, in contrast the rule based detection can do data cleaning in range of every type of recorded errors.